DOCUMENT RESUME

ED 062 401 TM 001 340

AUTHOR Garvin, Alfred D.
TITLE Confidence Weighting.
PUB DATE Apr 72
NOTE 12p.; Paper presented at the annual meeting of the American Educational Research Association (Chicago, Illinois, April 1972)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Achievement Tests; *Confidence Testing; *Guessing (Tests); Measurement Techniques; Multiple Choice Tests; *Objective Tests; Response Mode; Response Style (Tests); Statistical Analysis; Testing; Test Interpretation; *Test Reliability; *Weighted Scores
IDENTIFIERS Confidence Weighting

ABSTRACT

Various aspects of Confidence Weighting are examined. Variant of Confidence Weighting, its effect on test reliability, and the validity of Confidence Weighting are discussed. (DG)

ED 062401

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

Confidence Weighting

Alfred D. Garvin

University of Cincinnati

Inherent sources of unreliability in Objective Achievement Tests

Under conventional objective achievement testing procedure, the fact that a student has selected the "correct" response symbol for a given item says little about how much he actually knows about that item. All correct response symbols look alike, no matter how or why they were selected. One student might have been able to supply the correct response, without hesitation, to an open-ended question on the point involved. Another might not have been able to supply such a response but did recognize it at once when it was supplied. Still another might have just barely preferred the correct response over an incorrect alternative. Finally, another student may have selected this correct response quite by chance in a desperate flurry of random response selections during the final few seconds of the test. Thus, under conventional objective achievement testing procedure, response selections based on grossly disparate levels of relevant knowledge can receive the same score credit. A fortuitous "guess" receives full credit while relevant knowledge far beyond the minimum level required to divine the correct response cannot be manifested and, so, receives no extra credit.

The possibility of guessing and the necessity for dichotomous scoring are both inherent in conventional objective testing procedure. Guessing operates to inflate scores randomly at the lowest ability levels while

dichotomous scoring operates to truncate scores systematically at the highest ability levels. The resultant effects of these two factors are to reduce the range of scores and to introduce a random variable--chance. Both of these effects reduce test reliability. All of this was fully recognized from the beginning of objective testing but the potential utility of this test format inspired a search for procedural strategies to meliorate these inherent faults.

Potential Solutions

Corrections for Guessing. The most obvious source of unreliability was guessing and two so-called correction-for guessing strategies were developed. The well-known subtractive correction was intended to make guessing unprofitable; the less common additive correction was intended to make it unnecessary. Of course, it is impossible to "correct" for a random variable. If these two strategies have any effect on test reliability it is by inhibiting guessing on speeded tests and, even here, the effect will vary from testee to testee. However, most achievement tests are power tests and, as Gulliksen (1950) has pointed out, if every testee attempts every item, corrections for guessing have no effect on test reliability. In brief, these strategies are not the answer to all objective testing problems and may not be the answer to any.

Confidence Weighting. Ideally, an achievement test should permit the respondent to manifest all of the knowledge he has relevant to each item in the sample of items comprising the test. A dichotomously scored test merely counts the number of times he had "enough" knowledge. For every item on which he had more than "enough" knowledge, such scoring truncates the continuous underlying variable we are trying to measure. Confidence Weighting (CW) was designed to permit the testee to manifest his "extra" knowledge.

Confidence Weighting

Definition. CW is a special procedure for responding to objective test items, and for scoring such responses, wherein the respondent who is willing to indicate high confidence in his response selection is awarded a specified extra point credit if, indeed, he is right but he incurs a specified point penalty if, in fact, he is wrong. This option is exercised independently on each item. This procedure can be applied to any so-called objective-type item--true-false, multiple-choice, matching, or objectively scorable completion items. However, the empirical studies on CW reported in the literature have been confined to multiple-choice or true-false tests, with the latter somewhat more common.

Variants of CW. Oddly enough, the earliest studies on CW, dating from the mid-30's, employed the most elaborate procedures. In studies reported by Soderquist (1936) and Swineford (1938, 1941) involving both true-false and multiple-choice tests respondents had the option, on each item, of indicating any one of four levels of confidence in their response selection. Each level of confidence carried a different pair of score contingencies. The lowest level of confidence would yield 1 point if right, 0 if wrong; the next higher level would yield 2 or -4; the next level 3 or -6; and the highest level, 4 or -8. Different response symbols served to indicate the level of confidence the respondent wished to express in his response selection. The score contingencies specified for the lowest level of confidence, 1 or 0, will be recognized as those of conventional rights-only scoring. Dressel and Schmid (1953) offered the following pairs of score contingencies on a multiple-choice test: 1 or -1, 2 or -2, 3 or -3, and 4 or -4. Jacobs

(1968) compared the effects of two different bonus-penalty ratios, offering one group 1 or 0, 2 or -2, and 3 or -3, and the other group 1 or 0, 2 or -4, and 3 or -6. He found no signifiant differences between the risk-taking behavior patterns of these two groups. Much of the recent research done on CW has involved only two levels of confidence--none and some. In a series of studies using CW with true-false tests, Ebel (1965) offered contingencies of 1 or -1 and 2 or -2. It will be recognized that contingencies of 1 or -1 on a true-false test amount to a conventional subtractive correction for guessing. In addition, Ebel awarded .5 for each omission, which amounts to an additive correction for guessing. Thus, his procedure combined the features of CW and both forms of correction for guessing. Garvin (1969) reported an extensive study involving multiple-choice tests in which the only contingencies offered were 1 or 0 and 2 or -2. On certain of the tests involved, a quota was set such that a respondent could elect the "confident" option on no more than half the items. Of course, no minimum quota was ever set. Other patterns of contingencies and other special instructions have been employed in research on CW and in classroom testing practice but those described above are representative of the variants of CW in common use.

The Effect of CW on Test Reliability. Regardless of the particular procedure employed, the primary purpose of CW has been to improve test reliability. In almost every case report, it has done this, although the degree of improvement has varied widely from case to case. Moreover, widely disparate situational factors--test length, format, difficulty, and content, and respondent motivation--and, most important, disparate experimental methodologies, make it difficult to abstract generalizations from the studies cited here. Be all that as it may, the consensus of published reports on the effect of

CW on test reliability is that it does "work" to some degree. Further, the two contrary instances this writer has encountered serve to confirm his own theory about why it works when it works.

Since improvement in reliability has been the material dependent variable in this discussion thus far, it is necessary to provide a suitable metric for expressing this variable. Fortunately, two investigators in this field have independently arrived at the same metric for this purpose, although they have given it different names. Philip DuBois called it a Coefficient of Equivalent Length (CEL); Robert Ebel called it an Improvement Factor. Since the former is more explicitly descriptive, it will be used here.

It should be recognized that a test administered under CW procedure yields two score distributions--a rights-only or raw score distribution and a weighted score distribution that embodies the score bonuses and penalties due to CW contingencies. If the reliability coefficient of the raw scores (r_r) and of the weighted scores (r_w) are computed by any appropriate common algorithm, these may be compared to provide a measure of reliability improvement (or decrement) due to CW. The CEL compares these two reliabilities in a rearrangement of the Spearman-Brown Prophecy formula, viz.:

$$CEL = \frac{r_w(1 - r_r)}{r_r(1 - r_w)}$$

The CEL is interpreted as the factor by which a conventionally administered test would have to be lengthened (or shortened) to yield the reliability of the same test administered under CW procedure. A CEL > 1.0 indicates that CW has "worked"; a CEL ≤ 1.0 indicates that it has not. In this connection, it must

be remembered that the weighted scores on a test can be less reliable than the corresponding raw scores.

The earliest studies on CW merely reported the r_r and r_w obtained and let these test statistics "speak for themselves." However, it is possible to reconstruct a CEL for each of these studies and so compare these with later studies on a common basis. The CELs attained in the several studies cited herein are tabled below. The studies are listed chronologically; in the two multiple-experiment studies, CELs are listed in order of magnitude.

Hevner (1932)	1.72					
Soderquist (1936)	2.20					
Swineford (1938)	1.48					
Dressel and Schmid (1953)	1.16					
Ebel (1965)	1.00	1.07	1.19	1.48	1.72	1.84
Garvin (1969)	.96	1.19	1.38	1.64	1.84	

As previously noted, these results must be compared with caution in view of the disparate situations and methodologies involved. Nevertheless, the median CEL of 1.48 may be regarded as a reasonable expectation for the degree of reliability improvement to be expected in a typical test situation.

It will be noted that only one of the 15 CELs reported above is less than 1.0 and that only slightly so. Nevertheless, the possibility exists that the r_w of a given test would be much lower than its r_r. This raises both practical and ethical questions as to which set of scores should be used for various purposes. In anticipation of such a dilemma, this writer has made it a practice to advise his students that the more reliable of the two sets of scores would be used in determining grades. In over 80 testing events he has conducted under CW procedure, the weighted scores have been the more reliable in all but five cases.

The Effect of CW on Variation of the Standard Error of Measurement. The discussion thus far has concerned improving the global reliability of a test. Mollenkopf (1949) has shown that the standard error of measurement is not likely to be uniform over the range of scores in a distribution unless this distribution is normal. This is equivalent to saying that a test may be more reliable at one point in the score distribution than at another. Test results are generally used to partition testees at one or more points in the score distribution, e.g., assigning letter grades or selecting a high or low group for some special purpose. Accordingly, it may be more important to know how reliable our test is at the point or points where we are going to make our "cuts" than it is to know its "global" reliability.

The typical teacher-made, objective achievement test yields a negatively skewed raw score distribution. According to Mollenkopf's formulations, such a test is relatively more reliable in the extended, lower tail of the score distribution and relatively less reliable in the blunted, upper tail. If we are concerned only with the identification of some lowest group, this kind of test provides its highest reliability where it is most needed. If, however, a cut must be made somewhere in the upper end of the score distribution, the effective reliability of the test at this point is typically less than the global reliability of the test. It is not uncommon that a test is designed for one purpose and, sooner or later, its results are used for one or more other purposes. Against this possibility, the most desirable situation is that its reliability be high and uniform over the full range of scores. To attain this situation with the typical teacher-made, objective achievement test, the reliability of the upper end of the score distribution must be improved without simultaneously depressing the reliability of the lower end.

Garvin (1969) studied the effect of CW on variation of the error of measurement (over the range of test scores)--to quote the title of his dissertation. Eight sections of highly motivated, highly intelligent young men took each of five different tests (in trigonometry, spelling, and three aspects of English) under CW procedure. In 30 of these 40 section-by-test events, the variation of the error of measurement over the range of test scores was decreased by CW; when the eight sections were pooled within tests, this effect was found for every test. Thus, it would seem that CW does what it does--increase test reliability--where it is needed most--at the upper end of the score distribution.

The Validity of CW Procedure. Almost as soon as CW was developed, the construct validity of this procedure was challenged. Indeed, Swineford's first paper on the subject (1938) was entitled, "The Measurement of a Personality Trait." She contended that CW merely confounds achievement with an irrelevant personality trait--willingness to take risks (in a competitive academic setting). Jacobs (1968) substantially replicated her methodology and came to substantially the same conclusions. The implications of these conclusions are clear: two hypothetical students of equal "true" ability, one "confident" and one "diffident," would appear to be of unequal ability under CW procedure; boldness could eclipse wisdom.

Garvin's (1969) study hypothesized that, under conditions of earnest academic competition, relevant knowledge, confidence in one's knowledge, and willingness to manifest such confidence under the contingencies of extra credit vs score penalty are all highly and positively correlated. To test this hypothesis, he defined a subject's weighted score (X_w) minus

his raw score (X_r) as a measure of whatever it is that the CW procedure, itself, measures and he defined X_r, alone, as the a priori measure of whatever it was that the test, itself, measured. The product-moment correlation between this gain (or loss) due to CW and the raw scores, X_r, on the test was taken as a measure of concurrent validity for the CW procedure. Over the five tests involved in his study, these correlations ranged from +.49 to +.85 with a mean of +.69. It was concluded that CW measures more of the same thing that the test itself measures--relevant knowledge.

This rationale has been challenged on the grounds that the score component due to CW, $X_w - X_r$, is not independent of the raw score, X_r. This is quite true. The CW score compent for an individual is the resultant of his willingness to weight a given item and the probability of his being right when he does weight it, summed over all items. Willingness to weight and the probability of being right have been found to be highly correlated. The probability of being right, summed over items, is X_r and X_r is an a priori measure of relevant knowledge. Thus, the CW score component is related to X_r (and, so, to relevant knowledge) through the intervening variable, willingness to weight. If this were, in fact, a personality trait, uncorrelated with relevant knowledge, the high positive correlations found between the CW score component and raw scores would not have occurred. This expatiation of the writer's rationale for the empirical concurrent validity of CW does not settle this issue once and for all. It is simply one way to think about it. In the end, we must be pragmatic and look to the reliability coefficients involved. If we believe in these coefficients at all and in the dependence of validity on reliability, we must see some good in any testing procedure that quite consistently yields a higher reliability than conventional procedure would.

There is one more factor involved here that deserves consideration. CW, itself, may be said to have a kind of intrinsic validity. In certain content areas it can be just as important to know how confident a person is of his knowledge as it is to know how much knowledge he actually possesses. Consider, for example, the case of spelling. Imagine that two people spell a given word correctly on a spelling test. One was confident of his answer and would have weighted it under CW procedure; the other was not at all sure of his answer and would not have weighted it. Now, imagine, instead, that each of these two people was drafting a sentence in which this test word was appropriate. The first person would probably use the appropriate word; the second would probably use a less appropriate substitute that he was sure he could spell (or he would go to a dictionary, if one were available, and look it up--only to find that his "hunch" was right). It is of little practical value that he could, in fact, have spelled the original word correctly if he were forced to try. Imagine, next, that two other people spell this same word wrong on this test. One was quite unsure of his answer and would not have weighted it under CW procedure; the other was very sure that his answer was right and would have weighted it. Now, imagine, instead, that these two people are each drafting a sentence in which this word was appropriate. The first of these two people would probably substitute another word that he knew he could spell (or would consult a dictionary, if available); the second would probably go right ahead and make a glaring error--and never check it. Certainly, there is an important practical difference between the states of knowledge of the two people who both spelled the word correctly on the test and between the two who both spelled it wrong. A good teacher would do different things about each of these four people--if he knew that these four

different states of affairs existed. CW provides the teacher with a direct indication of each of these four states of affairs.

The importance of knowing the relationship between the state of a man's knowledge and his confidence therein and of doing different things about each combination of these variables was recognized long ago in an arabic maxim:

> He who knows not, and knows not that he knows not, is a fool. Shun him.
>
> He who knows not, and knows that he knows not, is simple. Teach him.
>
> He who knows, and knows not that he knows, is asleep. Waken him.
>
> He who knows, and knows that he knows, is wise. Follow him.

References

Dressel, P. L. & Schmid , J. Some modifications of the multiple-choice item. Educational and Psychological Measurement, 1953, 13, 574-595.

Ebel, R. L. Confidence weighting and test reliability. Journal of Educational Measurement, 1965, 2 (June), 49-57.

Garvin, A. D. The effect of confidence weighting on variation of the error of measurement. (Doctoral dissertation, University of Maryland) Ann Arbor, Mich.: University Microfilms, 1969. No. 69-7621.

Garvin, A. D. & Ralston, N. C. Improving the reliability of course pretests. Paper presented at the annual meeting of the National Council on Measurement in Education, Minneapolis, Minn., March 1970.

Gulliksen, H. Theory of mental tests. New York: John Wiley and Sons, 1950.

Jacobs, S. S. An empirical investigation of the relationship between selected aspects of personality and confidence-weighting behaviors. Unpublished Doctoral dissertation, University of Maryland, College Park, Maryland, 1968.

Mollenkopf, W. G. Variation of the standard error of measurement. Psychometrika, 1949, 14, 189-229.

Soderquist, H. O. A new method of weighting scores in a true-false test. Journal of Educational Research, 1936, 30, 290-292.

Swineford, F. Analysis of a personality trait. Journal of Educational Psychology, 1941, 29, 438-444.

Swineford, F. The measurement of a personality trait. Journal of Educational Psychology, 1938, 29, 289-292.